DOCUMENT RESUME

ED 377 222 TM 022 392

AUTHOR Oshima, T. C.; Davey, T. C.
TITLE Evaluation of Procedures for Linking Multidimensional Item Calibrations.
PUB DATE Apr 94
NOTE 24p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 5-7, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; Estimation (Mathematics); *Evaluation Methods; Matrices; Simulation; Statistical Distributions; *Test Items
IDENTIFIERS Calibration; Item Parameters; *Linking Metrics; *Multidimensionality (Tests); Variance (Statistical)

ABSTRACT

This paper evaluated multidimensional linking procedures with which multidimensional test data from two separate calibrations were put on a common scale. Data were simulated with known ability distributions varying on two factors which made linking necessary: mean vector differences and variance-covariance (v-c) matrix differences. After the calibrations of multidimensional item parameters, blocks of means from item parameter estimates were used to equate the two groups. The linking was effective for mean vector differences. The linking for v-c matrix differences was less effective, but encouraging. Suggestions for future research are provided. Four tables are attached. (Contains 6 references.) (Author)

Evaluation of Procedures for Linking Multidimensional Item Calibrations

T. C. Oshima
Georgia State University
and
T. C. Davey
American College Testing

Paper Presented at the Annual Meeting of National Council on Measurement in Education, New Orleans, April 1994

Running head: LINKING MULTIDIMENSIONAL ITEM CALIBRATIONS

Abstract

This paper evaluated multidimensional linking procedureS with which multidimensional test data from two separate calibrations were put on a common scale. Data were simulated with known ability distributions varying two factors which made linking necessary: mean vector differences and variance-covariance (v-c) matrix differences. After the calibrations of multidimensional item parameters, blocks of means from item parameter estimates were used to equate two groups. The linking was effective for mean vector differences. The linking for v-c matrix differences was less effective, but encouraging. Suggestions for future research are provided.

Evaluation of Procedures for Linking Multidimensional Item Calibrations

Unidimensional item response models have long been considered as somewhat unrealistic. Very few combinations of test items and examinee populations can be reasonably argued to produce truly unidimensional item response data. Fortunately, such arguments have become less important, as fast, convenient and inexpensive multidimensional item calibration computer programs have become more available (Fraser, 1987; Muthen, 1988; Wilson, Wood & Gibbons, 1991). That these programs have not been more widely used is perhaps due to a shortage of well established applications to practical testing problems. This is in contrast to the unidimensional case, for which reliable and intensely researched applications to test equating, test construction, item banking, and detection of differential item functioning have been developed (Lord, 1980).

A basic requirement of many practical applications is a way of linking items calibrated on different samples of examinees onto a common ability metric. While a variety of procedures have been proposed for the unidimensional case, extensions to the multidimensional case have not been carefully investigated. Davey and Kashima (Davey, 1991) have proposed a general framework for linking multidimensional calibrations. The purpose of this paper is to assess the performance of the linking procedures under this framework.

Linking multidimensional item calibrations

The compensatory (or linear) multidimensional item response model (McKinley & Reckase, 1983) expresses the probability of a correct response to the item $\underline{i}$ by an examinee with the ability vector θ_j as[1]:

$$Prob(u_{ij} = 1 \mid \boldsymbol{\theta}_j, \boldsymbol{a}_i, \boldsymbol{b}_i, c_i) = P_i(\theta_j) = c_i + (1 - c_i) L[\boldsymbol{a}_i^T(\boldsymbol{\theta}_j - \boldsymbol{b}_j)]$$

where L(•) is either the logistic or normal distribution function, the vectors $\underline{a}_i = \langle a_{i2}, \ldots, a_{ik}\rangle$ and $\underline{b}_i = \langle b_{i1}, b_{i2}, \ldots, b_{ik}\rangle$ characterize item discrimination and difficulty with respect to the $\underline{k}$ ability dimensions, and c_i gives the probability of an examinee with very low ability answering correctly by chance.

Because item and ability parameters enter the logistic function in the form $\underline{a}^T(\underline{\theta}-\underline{b})$, the vectors of discrimination parameters $\underline{a}$ can be linearly transformed (rotated) to $\underline{a}$ by premultiplying them by a nonsingular matrix $\mathbf{F}^T$, providing the ability and difficulty vectors, $\underline{\theta}$ and $\underline{b}$, are correspondingly premultiplied by the inverse, F^{-1}. Similarly, difficulty parameters can be translated by adding a vector of constants, $\underline{c}$, to each $\underline{b}$, provided these same constants are added to ability vectors as well. Thus, $\underline{\theta}^*=F^{-1}\underline{\theta} + \underline{c}$ and $\underline{b}^* = F^{-1}\underline{b} + \underline{c}$. Such transformations of the ability scale produce no net effect on item response surfaces.

Indeterminacy in the latent trait model is usually resolved by requiring obtained parameter estimates to satisfy some number of conditions or constraints. Imposing these constraints "identifies" the model and allows unique estimation of the remaining parameters. Unidimensional solutions are typically identified by setting the mean and variance of either the examinee ability or the item difficulty estimates to specified constants. For example, ability estimates may be scaled so as to have zero mean and unit variance. Multidimensional models require not only that the location and scale of each ability axis be fixed, but also that the orientation of the axes be specified. The simplest way of specifying the orientation of the ability axes is to set one or more item discrimination parameters to zero on a given dimension. However, more elaborate constraints are possible, and in fact desirable. For example, the mean of sets of discrimination parameters can be set to specified constants.

Estimating transformation parameters

The sum and substance of scale linking is finding rotation matrices, **F**, and translation vectors, $\underline{e}$, that take parameter estimates from separate calibrations to a common ability metric. For this to be possible, the separate calibrations must be based on common or randomly equivalent examinees, or include common items. The latter case is the principal focus here. The particular linking model considered regards one set of parameter estimates as a base that defines the ability space, while the second set of estimates is to be transformed to be consistent with that space.

Davey (1991) suggested estimating scaling parameters by simultaneously solving a set of scaling equations. Each scaling equation sets some function of the common item parameter estimates equal across calibrations by applying the proper choice of F and $\underline{e}$ to one set of estimates. The left hand side of each scaling equation is a function of the parameter estimates from the base calibration sample, while the right hand side is the same function of the other set of parameter estimates (or, more precisely, the transformed versions of these estimates). More formally, let $\underline{a}_1$, $\underline{b}_1$, and $\underline{\theta}_1$ denote the parameter estimates for the common items form the base calibration sample, while $\underline{a}_2$, $\underline{b}_2$, and $\underline{\theta}_2$ represent parameter estimates from the second calibration. The system of scaling equations then takes the form:

$$h_1(\underline{a}_1, \underline{b}_1, \underline{\theta}_1) = h_1(F^T\underline{a}_2, F^{-1}\underline{b}_2+\underline{e}, F^{-1}\underline{\theta}_2+\underline{e})$$

$$h_2(\underline{a}_1, \underline{b}_1, \underline{\theta}_1) = h_2(F^T\underline{a}_2, F^{-1}\underline{b}_2+\underline{e}, F^{-1}\underline{\theta}_2+\underline{e})$$

$$\vdots$$

$$h_q(\underline{a}_1, \underline{b}_1, \underline{\theta}_1) = h_q(F^T\underline{a}_2, F^{-1}\underline{b}_2+\underline{e}, F^{-1}\underline{\theta}_2+\underline{e})$$

where q is the number of elements of F and $\underline{e}$ to be estimated. The resulting, generally nonlinear, equations are solved simultaneously for the unknown elements of the rotation matrix F and the transformation vector $\underline{e}$.

While any properly structured set of scaling functions can serve to estimate scaling parameters[2], it is believed that more stable estimates will be obtained if the scaling equations themselves are stable functions of the item parameter estimates.

For example, equating the means of a large number of item parameter estimates is preferable to equating single estimates.

Methods

Design

This study concentrated on linking items calibrated under a two-dimensional model on samples from base and focal examinee populations. Using a compensatory multidimensional two-parameter logistic model, 40-item two-dimensional data sets were generated for the base and focal groups. Two factors were considered in this study to make scale linking necessary. The first factor, the difference between the mean vectors for the two groups, had three levels (no differences, small differences, and large differences). The base group always had a mean vector of [0, 0]. On the other hand, the focal group had three levels of the mean vector, (a) [0, 0], (b) [-.5, -.5], and (c) [-1, -1].

The second factor, the difference between the variance-covariance (v-c) matrix, also had three levels (no transformation, an orthogonal transformation, and an oblique transformation). The base group always had the v-c matrix of [1 .5, .5 1]. For the focal group, the v-c matrix was either (a) [1 .5, .5 1] which required no transformation, (b) [.8 .4, .4 .8] in which the variance was smaller but the correlation was the same as compared to the base group, requiring an orthogonal transformation, or (c) [1 .7,.7 1] in which only correlation differed, thus requiring an oblique transformation.

The two factors were completely crossed (3 x 3). Table 1 summarizes the nine conditions (C1 - C9). In each condition, examinees were drawn from either the base or the focal group, and responses to a common set of items were generated. Item parameters were then calibrated independently using NOHARM (Fraser, 1987) on both samples. Finally, the focal ability metric was linked to that of the base either by (a) a "poor" lining method or (b) a "good" linking method. These linking methods are described in the next section. In each condition, this process was repeated 20 times to produce distribution of linking and item parameter estimates, Consequently, 180 pairs (i.e., base and focal groups) of data sets were analyzed.

Insert Table 1 about here.

<u>Linking Methods</u>

Using a two-dimensional model, **F** was a 2 x 2 matrix and <u>e</u> was a two element vector. No restrictions was imposed on the structure of **F** and <u>e</u>, so a total of six scaling parameters were to be estimated, with six scaling equations required to do so. Two different methods for obtaining scaling equations were used. The first method, a "poor" linking method, equated only individual item parameter estimates across calibrations, and expected to yield poor estimates of the scaling parameters. More specifically, the six scaling equations were set using only two items of the test. The six equations are:

$$
\begin{aligned}
a_{1b2} &= a_{1f2}^* \\
a_{2b2} &= a_{2f2}^* \\
b_{1b2} &= b_{1f2}^* \\
b_{2b2} &= b_{2f2}^* \\
a_{1b3} &= a_{1f3}^* \\
a_{2b3} &= a_{2f3}^*
\end{aligned}
$$

where a_1 and a_2 stand for the elements of the a vector and b_1 and b_2 the elements of the b vector. The subscript 'b' or 'f' indicates the base or focal group, respectively. The last subscript indicates Item. Items 2 and 3 were used. The star * indicates transformed values. The choice of two specific items was arbitrary except avoiding Item 1 which always had a fixed a_2 from the NOHARM estimates.

The second method, a "good" linking method, equated the means of the blocks of item parameter estimates. In this study, all the items were used for the six scaling equations. The six equations are:

$$\sum_{i=1}^{20} a_{1bi} = \sum_{i=1}^{20} a_{1fi}*$$
$$\sum_{i=21}^{40} a_{1bi} = \sum_{i=21}^{40} a_{1fi}*$$
$$\sum_{i=1}^{20} a_{2bi} = \sum_{i=1}^{20} a_{2fi}*$$
$$\sum_{i=21}^{40} a_{2bi} = \sum_{i=21}^{40} a_{2fi}*$$
$$\sum_{i=1}^{40} b_{1bi} = \sum_{i=1}^{40} b_{1fi}*$$
$$\sum_{i=1}^{40} b_{2bi} = \sum_{i=1}^{40} b_{2fi}*$$

Again, the choice of the above six equations was arbitrary. The simultaneous equations for both linking methods were solved by using SAS/ETS in which the Newton method was used to solve equations.

Data Generation

Item parameters from a real test was used. The parameters are shown in Table 2. These parameters are estimates from a 1992 form of the ACT Assessment Mathematics test. Ability parameters (θ_1 and θ_2) were simulated from a random normal distribution. The mean vector varied from condition to condition as described earlier. The variance was also varied for some conditions. The appropriate correlated θ_1 and θ_2 for each condition were simulated by first generating two independent normally distributed pseudorandom variables z_1 and z_2 and then transforming them to θ_1 and θ_2 by weighted linear transformations. The weights were the elements of T^T, a matrix which satisfies $R = T^TT$, where R is the target correlation matrix. The sample size of each group was 1000.

Insert Table 2 about here.

Analysis

The linking procedures were compared at two levels. First, the mean linking parameter estimates from 20 replications were compared to true linking parameters. These true values were known since the population ability distributions were specified. Second, transformed item parameter estimates from the focal group were compared with true item parameters using root mean square error (RMSE). In addition, RMSE between the focal group item parameter estimates before transformation and true item parameters was obtained to establish the "no linking" baseline condition.

Results

Table 3 shows the comparison of true and estimated linking parameters. The elements of the matrix **F** are expressed as f1, f2, f3, and f4, and those of the vector $\underline{e}$ as e1 and e2. Clearly, the "poor" linking methods produced linking parameter estimates which are drastically different from true parameters. The estimates varied from replication to replication as indicated by rather large standard deviations. On the other hand, estimates from the "good" linking method are more consistent with true parameters. In addition, the smaller standard deviations over the replications suggest that the method produces fairly stable estimates.

Insert Table 3 about here.

The effects of the mean vector difference and the v-c matrix difference can be observed by comparing results from C1 through C9 for the "good" linking method. The mean vector difference are reflected in e1 and e2. The mean vector difference also affected the estimates of F. As the mean vector difference became large, estimates of F deviated from true parameters more (see C1, C4, and C7). As expected, the v-c matrix affected the estimates of F. However, the estimates are not quite consistent with true parameters for both orthogonal and oblique transformations.

Table 4 summarizes RMSEs between true and estimated item parameters after transformation. Reported are the mean and standard deviation of RMSE over the 20 replications. Again, it is obvious that the "poor" linking method produced item parameters which were very different from the true parameters. The mean RMSEs are large and the standard deviations are also large. Furthermore, the "poor" linking method was worse than no linking at all. The "good" linking method, on the other hand, appears to be an improvement over no linking except the orthogonal condition. The improvement was most obvious when there was a mean vector difference. Without linking, the deviation of d estimates can be serious. Throughout the conditions, the means and standard deviations of RMSE are fairly similar for the "good" linking method suggesting the effect of different mean vectors and v-c matrices are controlled somewhat via transformation.

Insert Table 4 about here.

Also noted in Tables 3 and 4 are the convergence problems. There were three cases (out of 360 cases) of convergence problems in the calibration process using NOHARM, and one case of convergence problems (out of 720 cases) in the linking process. The first occurred when examinee abilities were low (C8-C9), because one of the items (Item 39) was very difficult (d = -3.77) resulting in the p-value of zero. The second occurred for no obvious reasons. The simultaneous equations produced no solution for the particular case.

Discussion

The results of this study indicate that the degree to which item parameters from multiple calibrations of multidimensional test data are put on a common scale depended on how the scaling equations are selected to calculate the linking parameter estimates. When the equations are set as such that only a small portion of the test is used, the estimates for the linking parameters can be seriously distorted. In fact, in our example of the "poor" linking method where only two items were used for linking, the use of linking had more harm than merit. This phenomenon is understandable, because a small portion of items is hardly a representative of the entire test, and linking estimates from any deviant items within the small portion can distort the remaining item parameter estimates via an undesirable transformation.

The more important question is how to choose the scaling equations using the entire test. In this study, all the items were used and means of blocks were utilized for the solution to the scaling equations for the "good" linking method. However, there are many ways to choose the equations. Instead of using means for all the six equations, for example, standard deviations of item parameters can be used in some of the equations. There is a need for further studies in which various combinations of equations are compared to produce an optimal linking parameter estimates. In our results, the benefit of linking (i.e., the "good" linking method) was most obvious when there was a mean vector difference. When there was a v-c matrix difference, the results were not as straightforward as those from a mean vector difference. An oblique transformation seemed to produce item parameter estimates closer to true item parameters than an orthogonal transformation. In either case, however, the use of linking had no serious negative effect.

This study involved two stages of estimation. First, item parameters were estimated using NOHARM. Second, these item parameter estimates were used to estimate linking parameters. Therefore, any deviations of linking parameter estimates or item parameter estimates after transformation from true parameter values can be attributed to either the recovery ability of NOHARM or the estimation ability of the linking method, as well as the sampling errors. It appeared in our data using NOHARM that estimation was difficult in some correlated θs resulting some extreme item parameter estimates. These extreme estimates in turn affected the estimation of linking parameters.

The generalizability of the results is limited to the conditions examined in this study. Only one kind of multidimensional structure was used. Future studies need to focus on different multidimensional structures as well as different selections of scaling equations. In addition, there is a need for investigating the behavior of NOHARM in a comprehensive study.

References

Davey, T.C. (1991). Some issues in linking multidimensional item calibrations. Presented at the Office of Naval Research Contractors Meeting on Model-based Psychological Measurement, Princeton, NJ.

Fraser, C. (1987). NOHARM: An IBM PC Computer Program for Fitting Both Unidimensional and Multidimensional Normal Ogive Models of Latent Trait Theory [Computer Program]. Center for Behavioral Studies, The University of New England, Armidale, New South Wales, Australia.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Lawrence Erlbaum Associates.

Mckinley, R.L. & Reckase, M.D. (1983). An extension of the two-parameter logistic model to the multidimensional latent space. (Research Report ONR 83-2). Iowa City, IA. The American College Testing Program.

Muthen, B. (1988). LISCOMP: Analysis of linear structural equations with a comprehensive measurement model. [Computer Program]. Scientific Software, Inc.: Mooresville, IN.

Wilson, D., Wood, R. & Gibbons, R. (1991). TESTFACT: Test scoring, Item statistics, and item factor analysis [Computer program]. Scientific Software, Inc.: Mooresville, IN.

Notes

[1]Because the individual elements of $\underline{b}$ cannot be uniquely determined, the function argument is usually written as $\underline{a}^T\underline{\theta}+d$, where d is the product $\underline{a}^T\underline{b}$. However, the full parameterization is more convenient for purposes of scale linking. The transformation from reduced to full parameterization is arbitrary, in the sense that there are an infinite number of possible transformations. An especially intuitive transformation is the following, which distributes an item's difficulty across ability dimensions in proportion to the item's discrimination with respect to those dimensions:

$$\underline{b}_i = \underline{a}_i \frac{-d}{\underline{a}^T \underline{a}}$$

[2]The equations need only be independent and include as unknowns each of the scaling parameters.

[3]If, for example, **F** were required to be orthogonal, only three of its elements need be estimated, the fourth being determined. Other constraints on **F** and $\underline{e}$ can be considered.

Table 1. Distributions of the Focal Group

Condition	Mean θ_1	Mean θ_2	Variance	Covariance
C1	0.0	0.0	1.0	0.5
C2	0.0	0.0	0.8	0.4
C3	0.0	0.0	1.0	0.7
C4	-0.5	-0.5	1.0	0.5
C5	-0.5	-0.5	0.8	0.4
C6	-0.5	-0.5	1.0	0.7
C7	-1.0	-1.0	1.0	0.5
C8	-1.0	-1.0	0.8	0.4
C9	-1.0	-1.0	1.0	0.7

Note: For the reference group, [Mean θ_1, Mean θ_2] = [0.0, 0.0] and [Variance, Covariance] = [1.0, 0.5] throughout the conditions (C1 - C9).

Table 2. True Item Parameters

Item	a_1	a_2	d
1	2.3670	0.0000	2.4910
2	0.5770	0.3810	0.9950
3	0.6320	0.2650	0.6550
4	0.9900	0.4710	0.7640
5	0.5970	0.1830	0.2850
6	0.7700	0.6430	-0.0060
7	0.7530	0.4090	0.5680
8	1.6420	0.1470	1.2440
9	1.0400	0.5160	0.6090
10	1.2560	0.5140	0.2330
11	1.1710	0.1980	1.1100
12	1.2560	0.3880	0.9190
13	1.7110	0.4770	0.2370
14	0.6920	0.9140	-0.6760
15	0.5710	0.7210	-0.4360
16	0.3310	0.4270	-0.2750
17	2.0970	0.6940	0.5910
18	1.1900	1.1550	-0.9910
19	0.6320	0.3980	-0.2070
20	1.1120	1.3050	-0.6880
21	1.0200	1.1760	-0.0370
22	0.9560	1.2600	-0.4850
23	0.5970	0.8740	-0.5960
24	1.0100	0.4690	-0.1460
25	0.8330	0.7910	-1.4690
26	0.8140	0.7740	-1.0770
27	0.8690	0.8860	-0.9680
28	1.7110	1.7590	-0.0650
29	1.1170	1.1620	-0.8410
30	0.9280	1.3770	-1.1980
31	0.7940	1.3570	-0.9010
32	1.8670	1.5230	-1.3480
33	0.6020	0.4770	-0.6140
34	0.4420	0.4050	-0.9600
35	1.1520	2.1460	-1.8900
36	0.6250	0.7790	-0.8730
37	0.5260	0.9660	-1.3120
38	0.3060	0.9880	-1.5290
39	0.5710	2.2700	-3.7730
40	0.5640	0.4580	-0.8230
Mean	0.9672	0.8026	-0.3371
SD	0.4774	0.5265	1.0741

Table 3. Comparison of True and Estimated Linking Parameters (20 Replications in Each Condition)

(a) C1-C3

Condition	F E	True	Estimated Poor M	Poor SD	Good M	Good SD
C1	f1	1.00	-.05	(3.48)	1.02	(.14)
	f2	0.00	-.48	(4.75)	-.02	(.10)
	f3	0.00	2.38	(8.45)	-.05	(.22)
	f4	1.00	1.80	(9.88)	.96	(.20)
	e1	0.00	1.97	(12.84)	-.00	(.05)
	e2	0.00	-7.05	(43.36)	-.03	(.07)
C2	f1	0.89	.35	(1.67)	1.09	(.12)
	f2	0.00	.05	(1.44)	-.07	(.13)
	f3	0.00	1.57	(3.81)	.14	(.20)
	f4	0.89	.89	(3.12)	1.24	(.19)
	e1	0.00	-.47	(17.12)	-.01	(.05)
	e2	0.00	11.90	(48.03)	-.03	(.06)
C3*	f1	1.00	.92	(1.58)	1.03	(.15)
	f2	0.00	.26	(.80)	.07	(.24)
	f3	0.29	.12	(4.36)	-.24	(.24)
	f4	0.82	.20	(2.20)	.84	(.19)
	e1	0.00	-.46	(1.34)	.02	(.06)
	e2	0.00	1.38	(4.22)	-.04	(.06)

* C3 for the "good" linking method is based on 19 replications due to a convergence problem in one of the 20 replications in calculating linking parameters.

(b) C4-C6

Condition	F E	True	Estimated Poor M	Estimated Poor SD	Estimated Good M	Estimated Good SD
C4	f1	1.00	1.23	(1.66)	1.04	(.12)
	f2	0.00	-.09	(1.61)	-.00	(.10)
	f3	0.00	-.61	(4.38)	-.10	(.22)
	F4	1.00	1.33	(4.66)	.95	(.14)
	e1	-0.50	-.62	(.46)	-.50	(.11)
	e2	-0.50	-.38	(1.07)	-.46	(.19)
C5	f1	0.89	1.23	(1.44)	1.13	(.15)
	f2	0.00	.50	(1.87)	-.10	(.12)
	f3	0.00	-.16	(5.34)	.07	(.30)
	f4	0.89	-.42	(5.01)	1.36	(.24)
	e1	-0.50	-.75	(.63)	-.58	(.12)
	e2	-0.50	-.14	(1.30)	-.24	(.20)
C6	f1	1.00	3.76	(17.24)	1.06	(.14)
	f2	0.00	.88	(4.38)	.00	(.08)
	f3	0.29	-8.61	(51.15)	-.29	(.27)
	f4	0.82	-1.77	(12.72)	.88	(.20)
	e1	-0.50	-.87	(.55)	-.49	(.12)
	e2	-0.50	.25	(1.43)	-.60	(.31)

(c) C7-C9

Condition	F E	True	Estimated Poor M	Poor SD	Estimated Good M	Good SD
C7	f1	1.00	.34	(3.83)	.91	(.14)
	f2	0.00	.67	(3.43)	-.17	(.14)
	f3	0.00	3.33	(9.48)	.21	(.33)
	f4	1.00	-.74	(9.73)	1.31	(.29)
	e1	-1.00	-.86	(1.41)	-1.28	(.17)
	e2	-1.00	-1.45	(3.95)	-.35	(.40)
C8*	f1	0.89	-1.22	(8.70)	1.05	(.17)
	f2	0.00	-2.74	(13.80)	-.11	(.17)
	f3	0.00	8.15	(32.42)	.29	(.35)
	f4	0.89	11.96	(51.34)	1.56	(.29)
	e1	-1.00	-1.20	(.24)	-1.23	(.21)
	e2	-1.00	-.50	(.31)	-.20	(.39)
C9*	f1	1.00	.81	(.68)	.99	(.15)
	f2	0.00	.50	(.81)	-.08	(.14)
	f3	0.29	.39	(1.76)	-.15	(.33)
	f4	0.82	-.29	(2.08)	1.09	(.25)
	e1	-1.00	-1.34	(.63)	-1.11	(.23)
	e2	-1.00	-.09	(1.64)	-.75	(.50)

* C8 and C9 are based on 18 and 19 replications, respectively, due to a convergence problem in estimating item parameters using NOHARM.

Table 4. Root Mean Square Error Between True and Estimated Item Parameters after Linking (20 Replications in Each Condition)

		Poor			Good			No Linking		
		a_1	a_2	d	a_1	a_2	d	a_1	a_2	d
C1	M	4.04	3.32	5.41	.56	.41	.29	.59	.43	.30
	SD	7.07	6.31	10.27	.10	.13	.09	.16	.14	.10
C2	M	1.77	1.16	1.82	.55	.37	.24	.33	.27	.23
	SD	2.04	1.53	2.60	.13	.11	.08	.10	.10	.10
C3*	M	3.83	1.28	2.89	.61	.47	.39	.88	.57	.40
	SD	2.61	1.26	3.23	.12	.14	.14	.24	.20	.15
C4	M	2.21	1.88	2.11	.59	.47	.31	.65	.52.	1.45
	SD	2.79	3.91	2.60	.13	.13	.10	.21	.18	.23
C5	M	2.06	2.08	2.78	.54	.45	.27	.35	.32	1.22
	SD	3.23	2.57	4.25	.12	.07	.06	.08	.06	.09
C6	M	10.65	3.12	6.90	.66	.53	.40	.95	.68	1.67
	SD	32.94	4.98	15.57	.13	.12	.12	.30	.24	.33
C7	M	3.67	2.96	4.49	.58	.47	.30	.48	.40	2.37
	SD	6.18	6.86	8.32	.12	.09	.05	.15	.07	.21
C8*	M	5.78	8.51	8.78	.58	.49	.30	.31	.36	2.19
	SD	21.16	32.83	34.28	.13	.08	.03	.09	.07	.14
C9*	M	1.45	1.22	1.39	.64	.49	.36	.83	.51	2.75
	SD	.87	1.12	1.36	.12	.13	.09	.23	.18	.36

* C3 and C9 are based on 19 replications. C8 is based on 18 replications. See Table 2 for explanations.